

# The coordinate-based meta-analysis of neuroimaging data

Pantelis Samartsidis<sup>1</sup>, Silvia Montagna<sup>2</sup>, Thomas E. Nichols<sup>2</sup>, and Timothy D. Johnson<sup>3</sup>

<sup>1</sup>*MRC Biostatistics Unit, Cambridge CB2 0SR, U.K.*

<sup>2</sup>*Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.*

<sup>3</sup>*Biostatistics Department, University of Michigan, Ann Arbor, MI 48109, U.S.A*

## Abstract

Neuroimaging meta-analysis is an area of growing interest in statistics. The special characteristics of neuroimaging data render classical meta-analysis methods inapplicable and therefore new methods have been developed. We review existing methodologies, explaining the benefits and drawbacks of each. A demonstration on a real dataset of emotion studies is included. We discuss some still-open problems in the field to highlight the need for future research.

## 1 Introduction

*Functional Magnetic Resonance Imaging* (fMRI) has experienced a rapid growth over the past two decades and has lead to significant advances in our understanding of the human brain, including the differences in brain function between maternal and romantic love (Bartels and Zeki, 2004), the effect of alcohol while performing simulated driving (Calhoun and Pearlson, 2012), or the effect of doing nothing at all (Cole *et al.*, 2010). The availability of MRI scanners, inexpensive computational resources and accessible analysis software has made fMRI an ubiquitous tool in psychology, neurology and psychiatry, in addition to new areas like neuromarketing and neuroeconomics.

Nevertheless, there are a variety of factors that limit the interpretability of fMRI results. The principal limitation is the small sample sizes typically used, leading to individual studies suffer from low power and hence low reproducibility (Button *et al.*, 2013). Some other concerns include high prevalence of false positives (Wager *et al.*, 2007), poor reproducibility (Raemaekers *et al.*, 2007) and considerable heterogeneity in the analysis pipeline (Carp, 2012). As a result, it is unsurprising that the validity of fMRI is being challenged in both the scientific (Vul *et al.*, 2009) and popular (Shermer, 2008) literature (see (Farah, 2014) for an even-handed review).

Meta-analysis provides a way to address all of these limitations. Meta-analysis is the process of combining the results of independently conducted studies to increase power and obtain more reproducible findings than the original studies (Hedges and Olkin, 1985). Meta-analysis of functional neuroimaging data is an active field of research, whose growth is facilitated by the constantly increasing body of literature in fMRI along with the limitations of single experiments. The goal of this paper is to review recent advances in fMRI meta-analysis, evaluate existing methods, and highlight the open problems that need to be investigated. We consider a meta-analytic dataset analysed in a number of ways and make the data freely available to aid other researchers trying a hand at this intriguing area.

The remainder of this manuscript is organised as follows. In Section 2 we provide some background on neuroimaging and explain the special characteristics of neuroimaging studies that make meta-analysis in this setting challenging. In Section 3 we give a detailed description of the currently most popular methods for fMRI meta-analysis. In Section 4 we proceed to an evaluation of existing methods, which is motivated by application in a real dataset of emotion studies. Finally, we discuss some possible directions for research in Section 5.

## 2 Neuroimaging background

What follows is a very brief review of fMRI and the practical steps involved in a fMRI study. For a more detailed introduction, see Lindquist (2008) for review of fMRI for statisticians, or Kim and Ogawa (2012) for a detailed, technical review of the meaning of the fMRI signal; Huettel *et al.* (2009) provide an accessible textbook treatment, while Poldrack *et al.* (2011) give a practical, data-analysis-oriented perspective.

The objective of a single fMRI study is to identify the neural correlates of a physical, mental or perceptual process. When neurons in a region of the brain increase their firing rate, there is an increased demand for oxygen which is met by a localised increase in blood flow. The magnetic resonance signature, or susceptibility, of oxygenated and de-oxygenated blood differs, and thus a MRI scanner can capture changes in local oxygenation. This mechanism is known as the *Blood Oxygenation Level-Dependent* (BOLD) effect.

During an fMRI acquisition, participants lie flat in the scanner and are asked to perform a series of tasks, such as viewing images or reading texts, while the MRI scanner measures the BOLD signal. For each participant, the data takes the form of a time series of images, 3D snapshots of signal measurements all over the brain. The typical acquisition lasts 6-12 minutes, with data collected every 2 seconds, producing data on a grid with  $2\text{mm} \times 2\text{mm}$  spacing in-plane and 2mm-4mm slices, producing anywhere from 40,000 to over 100,000 voxels (volume elements) in the brain. Note that this is quite coarse spatial resolution, and separate, fine-resolution images (e.g.  $1\text{mm} \times 1\text{mm} \times 1\text{mm}$ ) are also taken to depict individual’s anatomy.

Before the raw data can be analysed, a series of preprocessing steps needs to be undertaken. These include motion correction, which accounts for movements during the acquisition, and spatial smoothing which increases the signal-to-noise ratio. To make data comparable across subjects, a crucial step is “spatial normalisation”, the process of warping all subjects to a standard brain template, or brain atlas. There are different atlases available, but essentially all authors use either the Talairach atlas (Talairach and Tournoux, 1988) or the MNI atlas (see Fig. 1).

After spatial normalisation, all subjects’ data exist in a common space. Specifically, we can assume that a given voxel corresponds to (roughly) the same region in all subjects’ brains. Statistical analysis then proceeds in a mass-univariate approach, fitting a model at each voxel independently of every other voxel. For every subject, time series regression models are used in each voxel, and then the subject-specific regression coefficients are subsequently analysed in a “second level” population model. The tradition in neuroimaging is for users to build their own General Linear Model and manually define contrasts to interrogate the model. Hence an individual effect (task) of interest is often referred to a “contrast”, meaning the estimated linear combination of parameter estimates that relates to the effect. The result is a 3D image of  $T$  statistics, one for each voxel in the brain, measuring the evidence against the null hypothesis of no effect. The  $T$  images are assessed either voxel-by-voxel,

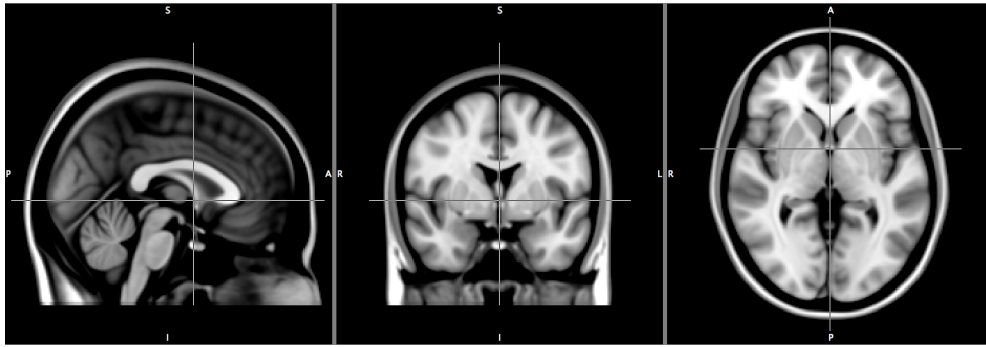


Figure 1: An average brain in MNI space. Note the directional labels at the edge of each panel: P for Posterior, A for Anterior, S for Superior, I for Inferior, L for Left and R for Right. The origin approximately corresponds to an anatomical structure known as the anterior commissure.

or by assessing the size of connected components, or *clusters*, after thresholding the  $T$  image at an arbitrary threshold. See [Friston \*et al.\* \(2002\)](#), [Mumford and Nichols \(2006\)](#) and [Mumford and Nichols \(2009\)](#) for a detailed review of different approaches for the statistical analysis of fMRI data.

An essential issue in the statistical analysis of fMRI data is *multiple testing*. A  $T$  statistic image can have 100,000 or more voxels in the brain, requiring 100,000 simultaneous tests for every contrast of interest. Under a global null hypothesis of no effect in any voxel, we therefore expect around 5,000 false positives using the classic significance level of  $\alpha = 0.05$ . In the early history of fMRI (roughly 1992-2002), arbitrary rule-of-thumb thresholding procedures were common, like a combination of an uncorrected voxel-wise  $\alpha = 0.001$  and cluster size threshold  $k \geq 10$  (only clusters of size 10 voxels or more). Thresholding methods that controlled the Familywise Error (FWE), the chance of one or more false positives, later became widespread using either Random Field Theory or permutation (see [Nichols and Hayasaka \(2003\)](#) for a review of FWE methods in neuroimaging). More recently, the *False Discovery Rate* (FDR), the expected proportion of false positives among positive findings ([Benjamini and Hochberg, 1995](#)) was introduced for the thresholding of  $T$  images ([Genovese \*et al.\*, 2002](#)).

In any other discipline of science, publishing the point estimate, standard error, test statistic and  $p$ -value for an effect would be considered best practice, if not just the minimal information to report. In neuroimaging, each of these quantities is 3D image, and sharing such large data files was considered impractical 20 years ago when fMRI was first developed. Yet even today there is general resistance towards sharing the full images. Instead, the only thing authors routinely report is the  $x, y, z$  atlas coordinates of activation peaks. Going forward we will call these coordinates the *foci* (singular *focus*). In other words, the results of an fMRI study are summarised in a list of foci. Based on author preference and software defaults, foci can either be *singly* reported that is one focus per significant region or *multiply* reported that is two or more per significant region.

## 2.1 Limitations of individual studies and meta-analysis

There are three aspects of fMRI experiments that challenge the utility of individual studies. Firstly, fMRI studies suffer from low power. The typical sample size of an fMRI study is small, and the majority of experiments involves far less than 20 participants ([Carp, 2012](#)). While power depends on the (unknown) true effect size, at least one empirical study supported the notion that fMRI  $n$ 's are too small. By sub-sampling from a large sample ( $n = 150$ ), [Thirion \*et al.\* \(2007\)](#) found that analyses with 20 or fewer subjects were poor approximations of the full 150-subject result. Further, Type I error rates are likely to be high, especially for older papers that did not use inference procedures

corrected for multiple testing. Using a survey of publications’ thresholding methods, [Wager et al. \(2007\)](#) estimated that 17% of all reported foci are false positives. Finally, neuroimaging studies suffer from low test-retest reliability. For example, when scanning a group of subjects twice, once and then 7 days later, [Raemaekers et al. \(2007\)](#) found intra-class correlations for BOLD fMRI activations ranged from 0 to 0.88.

Apart from these inherent limitations, the way fMRI studies are carried out also exhibits great heterogeneity. Each step of a neuroimaging study can be implemented in various ways and there is no standard way to present a stimulus, preprocess the data or construct the linear model for the BOLD response. As a result, there is only a partial agreement in how experiments are conducted. For example, in an analysis of 241 fMRI studies [Carp \(2012\)](#) observed 223 different analytical strategies. Results heavily depend on the type of analysis employed ([Button et al., 2013](#)), thus it is not uncommon to observe discrepancies in the outcomes of studies that investigate the same scientific question. Consequently, it is exceptionally hard to yield a conclusion. All reasons combined support the use of meta-analysis to account for these problems and draw more reliable inferences.

A well performed meta-analysis can tackle the aforementioned issues by modelling the observed heterogeneity between studies, combining the available information to increase power and ultimately separating the consistent findings from those that happened by chance. There exist two broad approaches for meta-analysis of neuroimaging studies: *image-based meta-analysis* (IBMA), if the full  $T$  statistic images are available, and *coordinate-based meta-analysis* (CBMA) if only foci are reported. IBMA proceeds by means of some common meta-analytic tools applied to each voxel of the images along with either FWER or FDR corrections for multiple testing (see [Hartung et al. \(2008\)](#) for an overview of conventional meta-analysis, and [Lazar et al. \(2002\)](#) for a review of IBMA methods). It is self-evident that the transition from full statistical images (100,000+ voxels) to the list of reported foci involves a heavy loss of information. In a comparative study, [Salimi-Khorshidi et al. \(2009\)](#) demonstrated the benefits of using IBMA over CBMA. However, the overwhelming majority of researchers rarely provide the full images, thus CBMA still constitutes the main approach for the meta-analysis of fMRI data. Thus, this review emphasises in widely used CBMA methods, which we now describe.

### 3 CBMA methods

The limitations of single experiments (see Section 2.1 for a discussion), along with the historical lack of data sharing, quickly presented researchers in the field of fMRI with a challenge. The standard meta-analytic tools used in other fields (see for example [Hartung et al. \(2008\)](#) for a fairly recent review) could not be applied to the coordinate data and hence there was a need for new methodologies. Early works mainly utilised exploratory data analysis and visualisation techniques to blend the results from different studies ([Fox et al., 1998](#)) and it was not until the early 2000’s that the first methods for CBMA were proposed ([Fox et al., 1997](#); [Turkeltaub et al., 2002](#); [Nielsen and Hansen, 2002](#); [Wager et al., 2003](#)). Since then, many new methods and modifications appeared in the neuroimaging ([Laird et al., 2005](#); [Wager et al., 2007](#); [Radua and Mataix-Cols, 2009](#); [Turkeltaub et al., 2012](#); [Caspers et al., 2014](#), to name a few) as well as the statistics ([Kang et al., 2011](#); [Yue et al., 2012](#); [Kang et al., 2014](#); [Montagna et al., 2016](#)) literature.

All of these methods share the same goal: to identify areas of the human brain that show consistent activation across studies. The different approaches broadly fall into two main categories: *kernel-based* and *model-based* methods. In what follows, we present the most widely used methods in both

categories. We start by setting the notation used throughout the chapter.

A typical CBMA dataset consists of a list of foci from  $I$  independent studies. Each study  $i$ ,  $i = 1, \dots, I$  comes with a set of 3-dimensional coordinates  $\mathbf{x}_{ik} \in \mathcal{B}$ , where  $\mathcal{B} \subset \mathbb{R}^3$  is the standard atlas space and  $k$  indexes the multiple foci for a particular study. Table 1 is part of a real dataset from a meta-analysis of emotion studies that will be analysed for the purposes of this review. In this example,  $\mathbf{x}_{52}$  would correspond to the second foci  $([-34, 52, 8]^T)$  in the fifth study (Baker 1997, emotion). Note that some of the studies (e.g. Damasio 2000, fear and Damasio 2000, anger) are obtained from the same experiment; we treat these studies as independent following the standard conventions in the field. Finally, we will denote as  $v = [v_x, v_y, v_z] \in \mathcal{B}$  the center location of a particular voxel in the brain atlas,  $v = 1, \dots, V$ .

Table 1: A subset of data from a meta-analysis study of emotions.

| Author   | Year | Emotion | X   | Y   | Z   | Participants |
|----------|------|---------|-----|-----|-----|--------------|
| Damasio  | 2000 | fear    | -10 | -62 | -17 | 23           |
|          |      |         | -1  | -66 | -1  | 23           |
|          |      |         | 34  | 3   | 32  | 23           |
| Damasio  | 2000 | anger   | -2  | -29 | -12 | 23           |
| Philips  | 2004 | disgust | 4   | -20 | 15  | 8            |
|          |      |         | 7   | -17 | 9   | 8            |
|          |      |         | 4   | -63 | 26  | 8            |
| Baker    | 1997 | sad     | 36  | 20  | -8  | 11           |
|          |      |         | -44 | 32  | -8  | 11           |
| Baker    | 1997 | happy   | -26 | 28  | 0   | 11           |
|          |      |         | -34 | 52  | 8   | 11           |
| Williams | 2005 | anger   | 7   | 31  | 28  | 13           |
|          |      |         | 7   | 28  | -7  | 13           |
| ...      | ...  | ...     | ... | ... | ... | ...          |

### 3.1 Kernel-based methods

The most widely used kernel based methods are the *multilevel kernel density analysis* (Wager *et al.*, 2007, MKDA), the *activation likelihood estimation* (Eickhoff *et al.*, 2012, ALE) and the *signed differential mapping* (Radua *et al.*, 2012, SDM). All these methods share the same rationale. Briefly, one starts by creating focus maps: that is, full brain images obtained through smoothing of reported activations with a spatial kernel. Obviously, there are as many focus maps as the total number of foci. Secondly, the focus maps corresponding to a particular study are combined to create the study-specific maps. These per-study images are subsequently combined into a single image that represents the evidence for consistent activation (clustering). Significance of these images is assessed with a Monte Carlo test under the null hypothesis of complete spatial randomness. We now discuss MKDA, ALE and SDM in detail.

#### 3.1.1 Multilevel kernel density analysis

First introduced by Wager *et al.* (2003), MKDA was modified to its current version by Wager *et al.* (2007). To obtain the focus maps,  $M_{ik}$ , one places a sphere of unit intensity and radius  $r$  centred at each focus:

$$M_{ik}(v) = \mathbf{1}_{\{d(v, \mathbf{x}_{ik}) \leq r\}} \quad (1)$$

where  $d(\cdot, \cdot)$  stands for the Euclidian distance. The study specific images,  $M_i$ , are then obtained by applying the maximum operator to the focus maps of the study. The procedure can be expressed by the following formula:

$$M_i(v) = \begin{cases} 1, & \exists k \text{ s.t. } d(v, \mathbf{x}_{ik}) \leq r \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

We call  $M_i$  the *comparison indicator maps*. A value of 1 means that there is activation within distance  $r$  of a given location. Wager *et al.* (2004) suggest giving  $r$  a value of 10 or 15mm. The MKDA statistic image  $m$  is given as a weighted combination of  $M_i$ :

$$m(v) = \frac{1}{\sum_i w_i} \sum_{i=1}^I w_i M_i(v) \quad (3)$$

The weights are usually chosen to be proportional to the number of participants in each study thus allowing for studies with larger sample size to contribute more to the value of the statistic. If the weights are all set to 1 then  $m(v)$  denotes the proportion of studies that reported activation within distance  $r$  to  $v$ . Large values of  $m(v)$  suggest systematic clustering of foci around its location.

The distribution of the MKDA statistic does not have a closed form and thus Monte Carlo testing is used to assess significance. Several synthetic datasets are created by uniformly drawing peak locations from  $\mathcal{B}$ , keeping the original number of foci fixed. The  $m$  statistic is calculated for these datasets and the maximum value is saved at each replicate. This produces a sample of the maximal statistic under the null hypothesis of random foci allocation. The sample is then used to obtain FWE corrected  $p$ -values (Kober *et al.*, 2008) as suggested by (Nichols and Holmes, 2002). Recently, Costafreda *et al.* (2009) derived a parametric significance test based on the properties of the spatial Poisson process. For applications of MKDA on real data see Etkin and Wager (2007) and Kober *et al.* (2008).

### 3.1.2 Activation likelihood estimation

The motivating idea behind ALE is to represent the uncertainty about the true location of a focus using a spatial Gaussian kernel (Turkeltaub *et al.*, 2002). Let  $L_{ik}$  be the map based on a single focus  $\mathbf{x}_{ik}$ ,

$$L_{ik}(v) = c\phi_3(v \mid \mathbf{x}_{ik}, \sigma_i^2 \mathbf{I}), \quad (4)$$

where  $\phi_3(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$  is the density of a three dimensional Gaussian distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$  evaluated at  $\mathbf{x} \in \mathbb{R}^3$ ,  $\mathbf{I}$  is the identity matrix, and  $c$  is the normalising constant. The Gaussian kernel used for ALE is analogous to the uniform kernel used for MKDA, but assigns higher values to the voxels closer to the foci. To help determine  $\sigma_i$ , Eickhoff *et al.* (2009) created a mapping between the number of participants in each study,  $n_i$ , and the standard deviation  $\sigma_i$ ; however, the mapping is based on an empirical study consisting of 21 subjects and may be unsuitable for experimental paradigms other than the one used by the authors.

Based on the Gaussian assumption,  $L_{ik}(v)$  represents the probability of  $v$  being the true location of  $\mathbf{x}_{ik}$ . These maps are combined into a *modelled activation map*,  $L_i(v)$ , giving the probability that the closest focus is truly located at  $v$  (Turkeltaub *et al.*, 2012):

$$L_i(v) = \max_k L_{ik}(v), \quad (5)$$



under the assumption that foci in study  $i$  are independent. The ALE statistic  $\ell$  is then computed as:

$$\ell(v) = 1 - \prod_{i=1}^I (1 - L_i(v)). \quad (6)$$

Expression 6 was originally adopted by Turkeltaub *et al.* (2002) to quantify the probability that at least one activation occurs in voxel  $v$ ; nevertheless,  $\ell$  should not be viewed as a probability distribution since that would require the foci of a study to be independent of one another.

The Monte Carlo significance test of ALE is equivalent but slightly different to the one of MKDA. In particular, Eickhoff *et al.* (2009) observe that one can directly derive null ALE values by recomputing  $\ell$  with random voxel location, i.e.:

$$\ell^* = 1 - \prod_i (1 - L_i(v^*)), \quad (7)$$

where  $L_i(v^*)$  are randomly selected voxels from the corresponding study map. Therefore, creating new datasets under the null hypothesis as in MKDA is not necessary. Recently, Eickhoff *et al.* (2012) showed that by using 7, it is possible to enumerate exhaustively all the possible outcomes, thus ensuring that the tail of the Monte Carlo distribution (where the inference is based) is better approximated. Thresholding to assess significance can be done by controlling the FDR or the FWE (Laird *et al.*, 2005) or by inferring on clusters rather individual voxels (Eickhoff *et al.*, 2012). ALE has been used for several analyses including Delvecchio *et al.* (2012) and Konova *et al.* (2013).

### 3.1.3 Signed differential mapping

SDM (Radua and Mataix-Cols, 2009) is a relatively new method that borrows several characteristics from both MKDA and ALE. The novelty of the method lies in incorporating the  $T$  statistic values (when available) from the original studies. To make this point clear, imagine that a study investigates brain activation caused by a given task; in some regions of the brain hyperactivation will be observed while in others there will be underactivation. In both cases, significant values of the  $T$  statistic will be recorded; these values will be large and positive in the first case and large and negative values in the second. This case is particularly interesting when difference in activation between tasks is being investigated.

Assume that  $T_{ik}$  is reported  $T$  value for the focus  $\mathbf{x}_{ik}$ . SDM will generate the focus maps  $S_{ik}$  as:

$$S_{ik}(v) = \text{sign}(T_{ik}) \exp\left(-\frac{d(v, \mathbf{x}_{ik})^2}{2\sigma^2}\right). \quad (8)$$

Observe that a Gaussian kernel is used, exactly as in ALE. Authors suggest using a standard deviation of approximately 10mm. The study maps are then:

$$S_i(v) = \begin{cases} -S_{\max}, & \sum_k S_{ik}(v) \leq -S_{\max} \\ \sum_k S_{ik}(v), & -S_{\max} \leq \sum_k S_{ik}(v) \leq S_{\max} \\ S_{\max}, & \sum_k S_{ik}(v) \geq S_{\max} \end{cases}, \quad (9)$$

where  $S_{\max} > 0$ . That is, the study map is obtained as the sum of the corresponding focus maps, but is forced to lie within the interval  $[-S_{\max}, S_{\max}]$  in the same way the MKDA study maps  $M_i$  are

given a maximum value of 1. Finally, the SDM statistic image,  $s$ , is calculated as the weighted mean of the study specific maps at each voxel:

$$s(v) = \frac{1}{\sum_i w_i} \sum_i w_i S_i(v). \quad (10)$$

Weights are once again proportional to number of participants in the study. Since the method averages both positive and negative findings, voxels that show contradicting results will not appear as significant. Inference is based on the same Monte Carlo scheme of MKDA and thresholding is done either by setting a highly conservative rejection point ( $p < 0.001$ ) or controlling the FDR (Radua and Mataix-Cols, 2009).

Recently, a new version of the algorithm was proposed, in which the authors used the  $T_{ik}$  values to reconstruct the original  $T$  statistic images. That way, it is possible to incorporate both CBMA and IBMA data in the same analysis. For more details, see Radua *et al.* (2012). The last contribution made on SDM lies in the use of anisotropic kernels in the analysis (Radua *et al.*, 2014). Anisotropy can be easily incorporated in MKDA and ALE but its superiority to the current practice of using isotropic kernels is only based on empirical findings and thus should be further investigated. Published work utilising SDM for the analyses includes Richlan *et al.* (2011) and Fusar-Poli (2012).

### 3.2 Model-based Methods

Recently there has been growing interest in the development of model based methodologies to address some of the limitations of kernel based methods. These methods use ideas from spatial statistics to develop stochastic models for the analysis of foci. Unfortunately the literature on model based methods is still very limited thus our review will be almost exhaustive. In particular, we will outline the *Bayesian hierarchical cluster process* model of Kang *et al.* (2011, BHICP), the *spatial binary regression* model of Yue *et al.* (2012), the *hierarchical Poisson/Gamma random field* model of Kang *et al.* (2014), and the *spatial Bayesian latent factor regression* model of Montagna *et al.* (2016). In all cases analyses are performed under the Bayesian paradigm and thus inferences are based on posterior distributions for each model's parameters.

Some of the methods reviewed here are build upon spatial point processes theory. Spatial point processes are random sets of points in the  $d$ -dimensional Euclidian space. A detailed description of the theory behind point processes is beyond the scope of this review, so we refer the reader to Møller and Waagepetersen (2004) and Illian *et al.* (2008) for details and applications. We now proceed to describe the details of model-based methods.

#### 3.2.1 A Bayesian hierarchical independent cluster process model (BHICP)

Kang *et al.* (2011) proposed a hierarchical model based on an independent cluster process to describe the mechanism generating the foci. The model is structured into 3 levels, of which the lowest level, level 1, contains the observations (foci), while higher levels describe the study and population structure respectively. The distinction between singly and multiply reported foci is incorporated into the model. In the outline of the model below, we occasionally suppress the  $k$  index so that  $\mathbf{x}_i$  is the set of foci reported in study  $i$ , i.e.  $\mathbf{x}_i = \bigcup_k \mathbf{x}_{ik}$ .

Figure 2 provides a graphical representation of the model. At level 1, we have foci (Fig. 2 bottom, coloured circles). We denote with  $\mathbf{X}_i$  for the underlying process generating the observations  $\mathbf{x}_i$  in



each study. As discussed in Section 2, we can have both multiply and singly reported foci. Thus,  $\mathbf{X}_i$  consists of two mechanisms, one generating the multiply reported foci (Fig. 2, red circles) and one that is giving the singly reported foci (Fig. 2 bottom, green circles):  $\mathbf{X}_i = \mathbf{X}_i^1 \cup \mathbf{X}_i^0$ .

Multiply reported foci  $\mathbf{X}_i^1$  can be viewed as an independent cluster process of points  $\mathbf{X}_{i\xi}^1$ , which are normally distributed around study activation centers  $\xi \in \mathbf{y}_i$  with covariance matrix  $\Psi_\xi$  (a random mark attached to every  $\xi \in \mathbf{y}_i$ ). That is,  $\mathbf{X}_i^1 = \bigcup_{\xi \in \mathbf{y}_i} \mathbf{X}_{i\xi}^1$  and for all  $\xi \in \mathbf{y}_i$ :

$$\mathbf{X}_{i\xi}^1 \sim \mathcal{N}(\xi, \Psi_\xi). \quad (11)$$

Singly reported foci,  $\mathbf{X}_i^0$ , come directly from the population activation centers  $\mathbf{z}$  as realisations of  $\mathbf{X}_{i\zeta}^0$ , which are normally distributed around population centers  $\zeta \in \mathbf{z}$  with covariance matrix  $\Sigma_\zeta$ :

$$\mathbf{X}_{i\zeta}^0 \sim \mathcal{N}(\zeta, \Sigma_\zeta). \quad (12)$$

To add more flexibility, the model allows for some singly reported foci to not cluster around any population center, say  $\mathbf{x}_{i\emptyset}$ . These foci are assumed to arise from a Poisson process  $\mathbf{X}_{i\emptyset}$  of constant intensity  $\epsilon_1$ :

$$\mathbf{X}_{i\emptyset} \mid \epsilon_1 \sim \mathcal{PP}(\mathcal{B}, \epsilon_1). \quad (13)$$

Overall,  $\mathbf{X}_i^0 = \left( \bigcup_{\zeta \in \mathbf{z}} \mathbf{X}_{i\zeta}^0 \right) \cup \mathbf{X}_{i\emptyset}$  are the singly reported foci of a study.

At level 2, we have the unobserved study activation centers  $\mathbf{y}_i$ , which are the locations around which the multiply reported foci of a study cluster. The  $\mathbf{y}_i$  are realisations of a point process  $\mathbf{Y}_i$  and may either cluster around the population centres  $\mathbf{z}$  (Fig. 2 middle, squares) or appear in random locations across the brain (Fig. 2 middle, triangles). To account for the former, clustered study centers,  $\mathbf{Y}_{i\zeta}$  are introduced as sets of points normally distributed around population centers  $\zeta \in \mathbf{z}$  with variance matrix  $\Sigma_\zeta$ . As for the latter, noise study centers are modelled as a homogenous Poisson process  $\mathbf{Y}_{i\emptyset}$  with intensity  $\epsilon_2$ . Overall,  $\mathbf{Y}_i = \left( \bigcup_{\zeta \in \mathbf{z}} \mathbf{Y}_{i\zeta} \right) \cup \mathbf{Y}_{i\emptyset}$ , where:

$$\mathbf{Y}_{i\zeta} \sim \mathcal{N}(\zeta, \Sigma_\zeta) \quad (14)$$

and

$$\mathbf{Y}_{i\emptyset} \mid \epsilon_2 \sim \mathcal{PP}(\mathcal{B}, \epsilon_2). \quad (15)$$

At the highest level (level 3), we have the population activation centres (Fig. 2 top, gray crosses). These are unobserved realisations  $\mathbf{z}$  of an a-priori homogenous Poisson process  $\mathbf{Z}$  of intensity  $\epsilon_3$ :

$$\mathbf{Z} \mid \epsilon_3 \sim \mathcal{PP}(\mathcal{B}, \epsilon_3) \quad (16)$$

Population activation centres are the locations around which study activation centers and singly reported foci scatter. As such, they can be viewed as locations in the brain where an overall population effect exists.

The BHICP can be viewed as a random effects model as it allows for both within-study and between-study variability. Samples from the posterior distributions are obtained via MCMC. Several interesting quantities can be inferred upon such as regions of consistent activations (through the distribution of populations centers), the uncertainty in the location of study centers around the population centers (through  $\Sigma_\zeta$ ) and the variability of the foci within studies (through  $\Psi_\xi$ ).

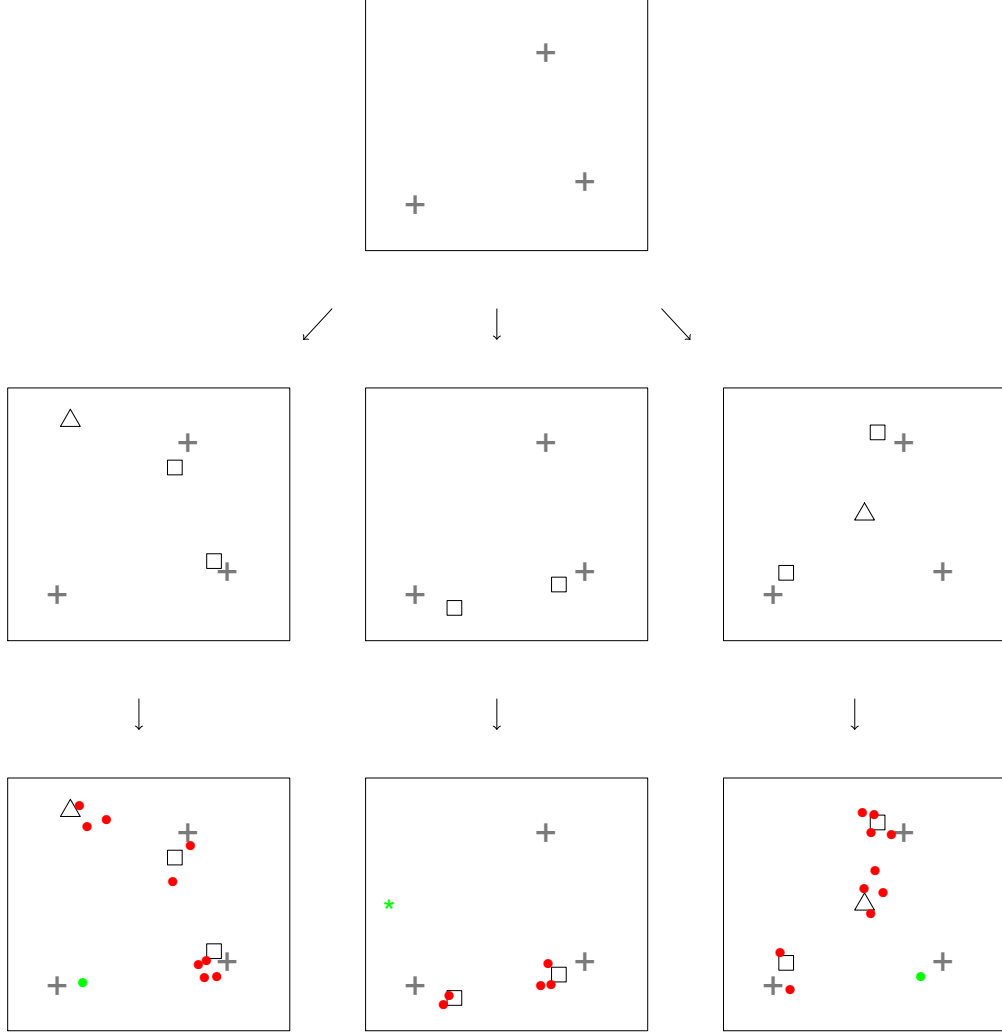


Figure 2: Realisation of the BHICP model for 3 studies. At level 1 (top) latent population centres (grey,  $\mathbf{z}$ ) lie. At level 2 (middle) we have centres of multiply reported foci (black). These come either directly from population centres (squares,  $\mathbf{y}_i$ ) or from background noise (triangles,  $\mathbf{y}_{i\emptyset}$ ). Level 1 (bottom) contains the data ( $\mathbf{x}_i$ ). These are multiply (red,  $\mathbf{x}_i^1$ ) or singly (green,  $\mathbf{x}_i^0$ ) reported foci. Singly reported foci come either directly from population centres (dots,  $\mathbf{x}_{i\xi}^0$ ) or from a background Poisson process (asterisks,  $\mathbf{x}_{i\emptyset}^0$ ).

### 3.2.2 A Bayesian nonparametric binary regression model

Yue *et al.* (2012) use spatial logistic regression for a meta-analysis of emotion studies. For study  $i$  and voxel  $v$ , let  $y_i(v)$  be the binary outcome defined as:

$$y_i(v) = \begin{cases} 1 & \text{at least one focus at voxel } v \\ 0 & \text{no foci at voxel } v \end{cases}. \quad (17)$$

Note that the binary study images  $\{y_i(v)\}_{v=1}^V$  are identical to the MKDA study maps  $M_i(v)$ . Logistic regression can be used to model the probability that a voxel is reported as a focus,  $p_i(v) = \mathbb{P}(y_i(v) = 1)$ . It is assumed that:

$$p_i(v) = H(z(v)), \quad (18)$$

where  $H(\cdot)$  is the link function. The authors use the standard probit and logit link functions.

Spatial correlation is induced through the prior on  $\{z(v)\}_{v=1}^V$ . In particular, we assume that the process  $z(v)$  is an adaptive Gaussian Markov random field (Yue and Speckman, 2010, aGMRF). The aGMRF model defines the conditional distribution of  $z(v)$  through a specific dependence with neighbouring voxels. A significant merit of the method is the inclusion of a local smoothness parameter  $\gamma(v)$  for the aGMRF. This allows the method to automatically choose the amount of smoothing required depending on the amount of information available.

Authors further introduce a process  $\psi_i(v)$ , an indicator of whether the outcome variable  $y_i(v)$  is miscoded; the case  $\psi_i(v) = 1$  can either refer to both false positives, voxels that were falsely found as activated, and false negatives, voxels that were not reported as foci even though they were activated. The process  $\psi(v)$  is not observed and hence is estimated along with the remaining model parameters.

Posterior probabilities of activation at each voxel are obtained through an auxiliary variable MCMC algorithm. Voxels with high posterior probabilities of being reported as foci are more likely to show an effect. A potential drawback of the method is that it can be currently applied only in two dimensions. In three dimensions, the value of one of the axes is held fixed, for example  $z = c$ , while the model is fitted for all available observations of the form  $\mathbf{x}_{ik} = [x1_{ik}, x2_{ik}, c]$ . Authors however, maintain that extending the model to three dimensions is possible.

### 3.2.3 A hierarchical Poisson/Gamma random field model (HPGRF)

A neuroimaging meta-analysis will typically consider several subtypes of tasks. For example, a meta-analysis of emotion may classify the studies according to experiments on “happiness”, “sadness”, “pain”, etc. Yet, the methods described previously are for a single homogeneous group of studies. Kang *et al.* (2014) propose a model that models each type of foci separately, allowing simultaneously for dependence between the  $J$  different types.

Let  $\mathbf{x}_{ij}$  be the set of foci reported by study  $i$  for task type  $j$ . Suppose that  $\mathbf{x}_{ij}$  are realisations of a Cox Process  $\mathbf{X}_j$  driven by a random intensity measure  $\Lambda_j(d\xi)$ . Conditional on  $\Lambda_j(d\xi)$ ,  $\mathbf{X}_j$  are Poisson processes on the brain  $\mathcal{B}$ :

$$\mathbf{X}_j \mid \Lambda_j(d\xi) \sim \mathcal{PP}(\mathcal{B}, \Lambda_j(d\xi)). \quad (19)$$

In other words, we have  $J$  underlying Cox processes, each one contributing a specific type of foci in some/all of the studies. The intensity measures  $\Lambda_j(d\xi)$  arise from a convolution of a finite kernel

measure  $\mathbf{K}_j(d\xi, \zeta)$  and a Gamma Random Field  $\mathbf{G}_j(d\zeta)$ :

$$\Lambda_j(d\xi) = \int_{\mathcal{B}} \mathbf{K}_j(d\xi, \zeta) \mathbf{G}_j(d\zeta). \quad (20)$$

The model arising from (19)-(20) is similar to the Poisson/Gamma random field model of [Wolpert and Ickstadt \(1998\)](#), who first introduced the idea of convolving a Gamma random field with a Poisson process. To introduce dependence between the different tasks, it is assumed that  $\mathbf{G}_j(d\zeta)$  are independent realisations of a Gamma random field with common shape measure  $\mathbf{G}_0(d\zeta)$  and inverse scale parameter  $\beta$ :

$$\mathbf{G}_j(d\zeta) \sim \mathcal{GRF}(\mathbf{G}_0(d\zeta), \beta). \quad (21)$$

Again,  $\mathbf{G}_0(d\zeta)$  is a Gamma random field:

$$\mathbf{G}_0(d\zeta) \sim \mathcal{GRF}(\alpha(d\zeta), \beta_0). \quad (22)$$

An MCMC scheme is used for posterior computation. The HPGRF model allows for the detection of overall effects based on the posterior intensity  $\mathbf{G}_0(d\zeta)$  or task-specific effects based on  $\Lambda_j(d\xi)$ . Inference on types with fewer observations can be done by borrowing information from the remaining types through correlation under the common base intensity  $\mathbf{G}_0(d\zeta)$ . A significant benefit of the model is that it requires the specification of very few hyperparameters.

### 3.2.4 A spatial Bayesian latent factor regression model

In a recent contribution, [Montagna et al. \(2016\)](#) generalise the model proposed in [Montagna et al. \(2012\)](#) to the case where observations are spatial point patterns from different contrasts, hereafter called studies. Following the convention of existing neuroimaging CBMA, the studies are treated as independent.

Consider independent spatial point patterns arising from  $n$  studies,  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . The authors regard  $\mathbf{x}_i$  as a Cox process  $\mathbf{X}_i$  driven by a non-negative random intensity function  $\mu_i$  defined on a common brain template  $\mathcal{B}$ :

$$\mathbf{X}_i \mid \mu_i \sim \mathcal{PP}(\mathcal{B}, \mu_i). \quad (23)$$

[Montagna et al. \(2016\)](#) consider a functional representation for the (log) intensity function and write  $\log \mu_i$  in terms of a collection of basis functions:

$$\log \mu_i(\boldsymbol{\nu}) = \sum_{m=1}^p \theta_{im} b_m(\boldsymbol{\nu}) = \mathbf{b}(\boldsymbol{\nu})^\top \boldsymbol{\theta}_i. \quad (24)$$

This specification implies that  $\log \mu_i$  belongs to the span of a (fixed) set of basis functions,  $\{b_m(\cdot)\}_{m=1}^p$ , with  $\boldsymbol{\theta}_i$  denoting a vector of study-specific coefficients. For a discussion on the choice of the bases (e.g., B-splines, Gaussian kernels, etc.) and their number  $p$  we refer to [Montagna et al. \(2016\)](#). A low dimensional representation of  $\log \mu_i$  is achieved by placing a sparse latent factor model ([Arminger and Muthén, 1998](#)) on the basis coefficients:

$$\boldsymbol{\theta}_i = \mathbf{\Lambda} \boldsymbol{\eta}_i + \boldsymbol{\zeta}_i, \quad \text{with} \quad \boldsymbol{\zeta}_i \sim N_p(0, \boldsymbol{\Sigma}) \quad (25)$$

where  $\boldsymbol{\theta}_i = [\theta_{i1}, \dots, \theta_{ip}]^\top$ ,  $\mathbf{\Lambda}$  is a  $p \times k$  factor loading matrix with  $k \ll p$ ,  $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{ik})^\top$  is a

vector of latent factors for study  $i$ , and  $\boldsymbol{\zeta}_i = (\zeta_{i1}, \dots, \zeta_{ip})^\top$  is a residual vector that is independent with the other variables in the model and is normally distributed with mean zero and diagonal covariance matrix  $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ .

Two attractive features of this approach are the ability to accommodate covariate information (*meta-regression*) and perform *reverse inference*. Both goals are achieved by putting the low dimensional vectors of latent factors  $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$  in any flexible joint model with other variables of interest. For example, information from covariates  $\mathbf{Z}_i$  can be incorporated through a simple linear model:

$$\boldsymbol{\eta}_i = \boldsymbol{\beta}^\top \mathbf{Z}_i + \boldsymbol{\Delta}_i, \quad \text{with} \quad \boldsymbol{\Delta}_i \sim N_k(0, \mathbf{I}), \quad (26)$$

where  $\boldsymbol{\beta}$  is a  $r \times k$  matrix of unknown coefficients, and  $r$  denotes the dimension of  $\mathbf{Z}_i$ .

Finally, reverse inference refers to inferring which cognitive process or task generated an observed activation in a certain brain region. In mathematical terms, it corresponds to estimating  $\Pr[\text{Task} \mid \text{Activation}]$ . Suppose for simplicity that the meta-analysis dataset consists of studies that can be categorized as either type A (e.g., happy) or type B (e.g., sad). Let  $y_i$  denote the study type, with

$$y_i = \begin{cases} 1 & \text{if study } i \text{ is type A} \\ 0 & \text{if study } i \text{ is type B.} \end{cases}$$

The interest is in estimating the probability that newly observed point pattern data arose from either a type A or B experiment. Because the study type can be represented as a binary response, the authors build a probit model for study type and predict the posterior probability that a new point pattern data arose from either type. Specifically, they model  $p_{y_i} = \Pr(y_i = 1 \mid \alpha, \boldsymbol{\gamma}, \boldsymbol{\eta}_i) = \Phi(\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\eta}_i)$ , where  $\Phi(\cdot)$  denotes the standard normal distribution function. Parameter  $\alpha$  can be interpreted as the baseline probability that study  $i$  is of type A, and  $\boldsymbol{\gamma}^\top \boldsymbol{\eta}_i$  accounts for study-specific random deviations. Notice that the latent factors  $\boldsymbol{\eta}_i$  are used as a vehicle to link the random intensities (thus, the foci) to the study-type. We remark that the probit model can be easily replaced by an appropriate predictive model for categorical, nominal, or continuous study features.

## 4 Evaluation of existing methods

One of the aims of this dissertation is to evaluate CBMA methods. A head-to-head comparison of existing methodologies is unfeasible, because the statistics described earlier have very different interpretations. Instead, we examine some characteristics of CBMA methods that show the drawbacks and merits of each. In what follows, we focus on the comparison between kernel-based and model-based methods. In Section 4.1 we conduct a series of simulations to study the sensitivity properties of the ALE algorithm that we think characterise other kernel-based methods as well. In Section 4.2, we apply the methods for which available software exist on a real dataset and compare the outputs. Finally, in Section 4.3 we proceed to a discussion.

### 4.1 ALE simulation study

Even though kernel-based methods have been extensively used for the analysis of neuroimaging data, their power properties have not been investigated on synthetic datasets. We perform a simulation study to assess the power properties of the ALE method. In particular, we want to assess how the power of the algorithm evolves with respect to the number of studies in the meta-analysis and whether

the method is robust to the inclusion of low quality studies. We choose ALE for three main reasons. Firstly, ALE is currently the most broadly used method for CBMA (based on a PubMed search for ALE, MKDA and SDM). Secondly, a recent review of kernel-based methods (Radua and Mataix-Cols, 2012) reported that the three kernel-based methods provide qualitatively similar results, thus we expect that our findings are indicative of MKDA and SDM methods as well. Finally, we strongly believe that the current version of ALE (Eickhoff *et al.*, 2012) provides the best approximation to the Monte Carlo test null distribution upon which inference is based.

We create meta-analytic datasets based on the following setup. Each simulated dataset consists of  $I$  studies; of these,  $Ip$  are valid while the rest  $I(1 - p)$  are noise,  $0 \leq p \leq 1$ . For the valid studies, we assume there exist 8 population centers around which foci cluster. A valid study detects each population center independently with probability 0.8. Conditional on detection, a study will report a singly reported focus with probability 0.4, two multiply reported foci with probability 0.35 or three multiply reported foci with probability 0.25. The foci are drawn from a three dimensional Gaussian distribution centered at the corresponding population center. As for the noise studies, we simply sample foci uniformly from the brain mask. The expected number of foci for both valid and noise studies is set low to 13, in order to resemble the sparsity of points that we observe in real applications.

We use study numbers of  $I$  of 20, 40, 60, 80, 100 and 120. For a given  $I$  we successively set  $p = 0, 0.05, 0.10, 0.15, \dots, 0.95, 1$ . For each distinct combination of  $I$  and  $p$  we create  $B = 1000$  datasets as described above, and apply the ALE algorithm (Eickhoff *et al.*, 2012) to each dataset. We use an  $\alpha = 0.05$  FDR corrected threshold to assess significance of the ALE statistic images. The following power-related quantities are recorded: 1) the probability that at least one of the 8 population centers is detected; 2) the probability all 8 centers are detected; 3) the mean number of centers detected in 1000 runs; 4) the mean voxel-wise true positive rate, where “truly” active voxels are defined by the 95% probability spheres around the population centers.

Our findings are summarised in Figure 3 where quantities 1 – 4 are plotted against the proportion of valid studies. One can observe that all 4 power measures increase monotonically to their maximal values of 1, 1, 8 and 1, respectively, as the number of studies grows. For a given proportion of valid studies, we observe that the bigger the total sample size is, the higher the power. For a fixed sample size, the power increases with the proportion of valid studies. Therefore, ALE is a consistent test. In Figure 4 we plot quantities 1 – 4 versus the total number of valid studies, that is,  $Ip$  instead of  $p$ . We see that the curves for different  $I$  tend to coincide. This is a key robustness property of the ALE algorithm: that is, adding pure-noise studies does not degrade power detection.

## 4.2 Analysis of a real dataset

In this section, we perform a meta-analysis of emotion studies that will facilitate the discussion of the next section. The dataset consists of 164 experiments conducted between 1993 and 2005. Eight emotion types appear in the dataset: affective, anger, disgust, fear, happy, mixed, sad and surprise. A total number of 2478 foci is reported with a mean value of foci per study close to 6. The goal of the analysis is to find regions of consistent activation across emotions. Due to the lack of software availability we only apply MKDA<sup>1</sup>, ALE<sup>2</sup>, SDM<sup>3</sup> and the BHICP<sup>4</sup>. Since those methods can not account for different task types we treat different emotions within an experiment as independent; this

<sup>1</sup><http://wagerlab.colorado.edu/files/tools/meta-analysis.html>

<sup>2</sup><http://www.brainmap.org/ale/>

<sup>3</sup><http://www.sdmproject.com/software/>

<sup>4</sup><http://www-personal.umich.edu/~jiankang/software.html>



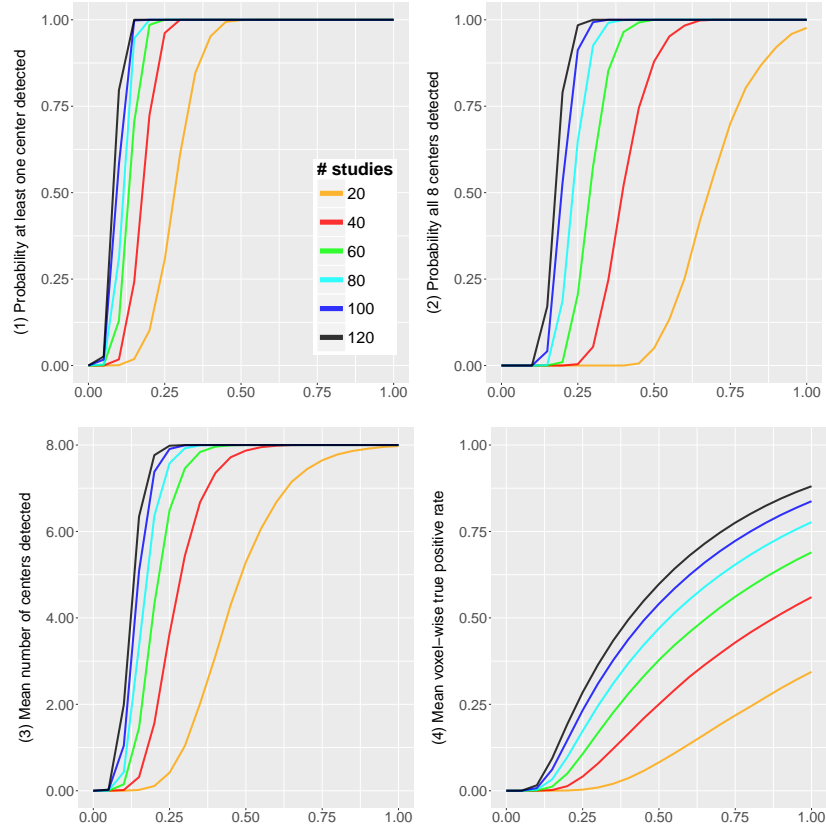


Figure 3: Results of the simulation study. Power properties of the ALE algorithm are plotted against the proportion of valid studies  $p$ . Top left: probability at least one center detected. Top right: probability all 8 centers detected. Bottom left: mean number of centers detected. Bottom right: mean voxel-wise true positive rate.

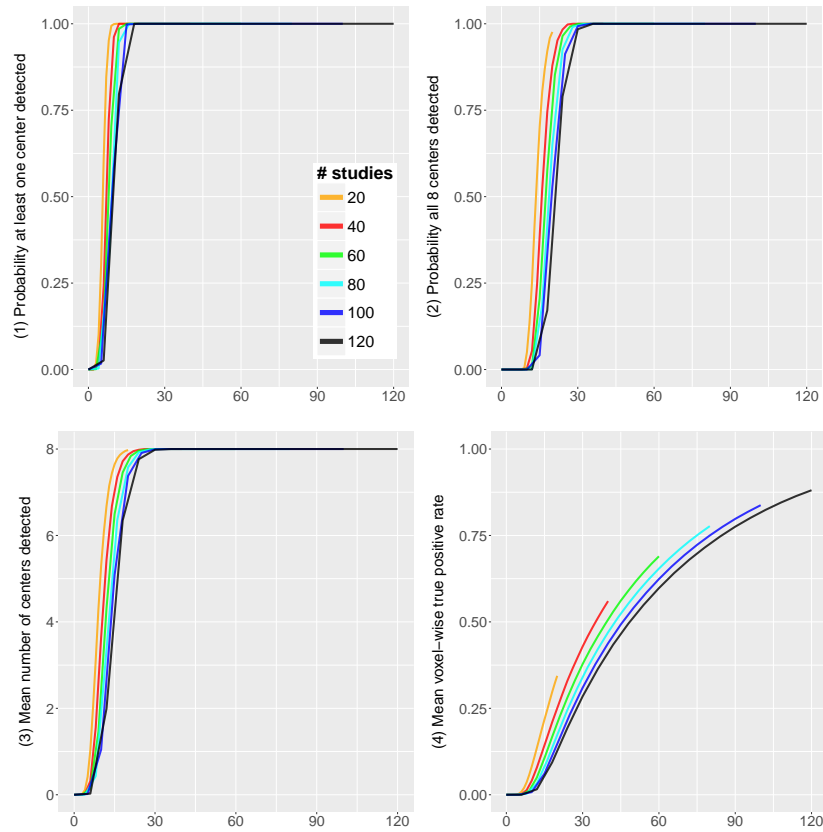


Figure 4: Results of the simulation study. Power properties of the ALE algorithm are plotted against the total number of valid studies  $I_p$ . Top left: probability at least one center detected. Top right: probability all 8 centers detected. Bottom left: mean number of centers detected. Bottom right: mean voxel-wise true positive rate.

results into a total sample size of  $I = 437$  studies (contrasts). The same dataset was analysed by [Kober \*et al.\* \(2008\)](#), [Kang \*et al.\* \(2011\)](#) and [Yue \*et al.\* \(2012\)](#).

The simulation parameters are set as following. For MKDA, we use a kernel size of  $r = 10\text{mm}$ , which is also the software default. A total of 10,000 Monte Carlo datasets are generated under the null hypothesis and used to threshold the MKDA statistic image  $m(v)$  at  $\alpha = 0.05$ , FWE corrected. ALE automatically assigns a kernel size for each study based on the total number of participants and uses the method of [Eickhoff \*et al.\* \(2012\)](#) to calculate the distribution of the statistic under the null hypothesis. The significance of the statistic image  $\ell(v)$  is assessed with an FDR corrected  $\alpha = 0.05$  threshold. For SDM we use an isotropic kernel of 20mm since it is the software default and do 500 Monte Carlo randomisations. For the BHICP we run the MCMC for 120,000 iterations saving once every 100 iterations. This results to a total sample size of 1200 posterior draws, of which we discard the first 200 as burnin. We use the same hyperparameter values as in [Kang \*et al.\* \(2011\)](#). We now summarise the results.

Figure 5 shows statistic images obtained from the four methods considered, conditional on several values of the  $z$  dimension. Note that for the BHICP we show only the study activation process intensity function. We see that all of the methods provide qualitatively similar results. More specifically, the regions of the brain that are mostly engaged in emotion processing are the right and left amygdala (Fig. 5, top and middle row). This finding is consistent with previous analyses of the same dataset ([Kober \*et al.\*, 2008](#); [Kang \*et al.\*, 2011](#); [Yue \*et al.\*, 2012](#)) as well as results of previous studies ([Phelps and LeDoux, 2005](#); [Costafreda \*et al.\*, 2008](#)). Other regions where large statistic values are observed are the right and left cerebral cortex (Fig. 5, bottom row).

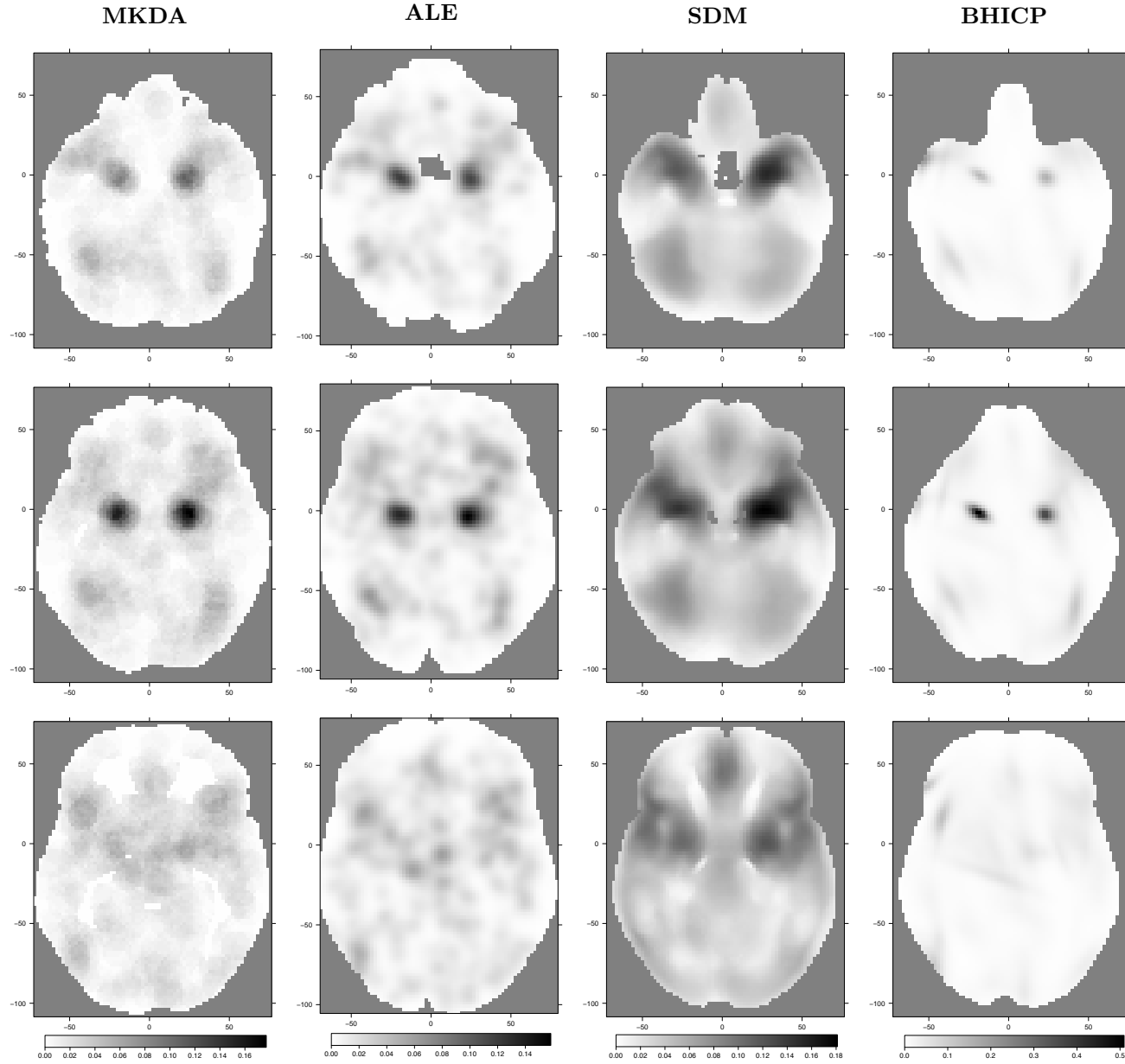


Figure 5: Qualitative comparison of methods in regions of high clustering of foci. Column 1 is MKDA, column 2 is ALE, column 3 is SDM and column 4 is the study activation center intensity. Rows 1-3 correspond to axial slice  $z = -22$ ,  $z = -16$  and  $z = -2$  respectively.

### 4.3 Discussion

The meta-analysis of emotions suggests that the results obtained from model-based methods are qualitatively similar to those obtained from kernel-based methods. However, it is difficult to compare these results quantitatively as they have very different interpretations. For example, the value of the MKDA map at a particular voxel denotes the weighted proportion of studies that report a focus within a pre-specified distance of that voxel. For model-based approaches based on point processes, instead, the integrated intensity over the brain (or over any ROI) is interpreted as the expected number of foci over the brain or ROI.

In terms of computing time, kernel-based methods outperform model-based methods (with the exception of SDM). For the emotion dataset used in Section 4.2, ALE required approximately 15 minutes to run whereas MKDA required around 3 hours for 10.000 Monte Carlo replications. On the contrary, the BHICP model took roughly 16 hours for 120.000 MCMC iterations, Kang *et al.* (2014) needed 20 hours to complete the analysis and it is yet not possible to run the spatial binary regression model on the full brain.

Additional considerations can be drawn regarding the implementation of these approaches. On the one hand, software for kernel-based methods can be applied to any dataset and will produce a pair of brain images, one with the value of the statistic at each and every voxel and one containing the corresponding  $p$ -values. By directly comparing these two images, one can easily identify significant voxels. On the other hand, it is not straightforward to implement an MCMC scheme for the Bayesian model-based methods discussed in Section 3.2. Prior specifications that are suitable for one dataset may be completely inappropriate for another. Further, it is not possible to know in advance how many iterations are required for the MCMC algorithm to converge, and convergence needs to be assessed as well. From a purely computational perspective, therefore, it may not seem appealing to a neuroimaging practitioner to adopt model-based methods, and this might explain why kernel-based methods are generally preferred.

Kernel-based methods, however, suffer from some serious limitations. First, kernel-based methods are based on a *mass univariate approach* (MUA) that lacks an explicit spatial approach to the modeling of the foci. Specifically, MUA consists of fitting a model at each voxel independently of every other voxel. Even though the activation of nearby voxels is correlated, estimation with the MUA ignores the spatial correlation, but inference later accounts for it when random field theory of permutation procedures define a threshold for significant activation. By failing to capture the spatial nature of the data, kernel-based methods do not provide an accurate representation of the true data generating mechanism and can not jointly characterize randomness of the number and locations of activations within each study. Further, there is no measure of uncertainty associated with the effect estimate, and conclusions could be misleading. For example, in Section 4.1 we found that power properties of ALE do not degrade with the inclusion of poor quality studies (see Fig. 4). Since inferences remain unchanged, it is not possible to distinguish between cases with strong signal (few poor quality studies) and weak signal (many poor quality studies). Note that this is a fixed effects model property, where a small proportion of the data drives the inference. Model-based methods tackle this problem by providing standard errors obtained for any parameter of interest directly from the posterior distribution; when the signal is strong there will be small variability in the posterior estimates whereas when signal is weak uncertainty will be higher. Finally, there is no adequate justification behind the choice of the kernel size parameter of kernel-based methods ( $r$  for MKDA and  $\sigma$  for ALE, SDM). Typically, its value is specified based on previous studies rather than being estimated from the data, and it remains

constant across the brain regardless of the amount of smoothing required in each region. However, a bad choice for the kernel size can potentially affect the results. For example, in the third column of Figure 5 we observe bigger clusters resulting from a choice for a larger kernel size than that used for MKDA and ALE. Yue *et al.* (2012) automatically choose the required amount of smoothing by introducing an extra smoothness parameters in their GMRF.

Several other quantities of interest can be obtained from model-based methods. For the BHICP model it is possible to derive  $(1 - \alpha)\%$  credible ellipses for both population and study activation centers, thus returning an estimate of within-study and between study-variability as in a random effects meta-analysis model. By introducing the latent process  $\psi(v)$ , Yue *et al.* (2012) estimate the probability of a voxel being miscoded. In the HPGRF model, the authors provide correlation estimates between the different emotions. All these quantities can not be obtained by any of the kernel-based methods.

Finally, the Bayesian framework upon which model-based methods are built facilitates the construction of predictive distributions over new studies. This helps producing the so-called *reverse inferences* (Poldrack, 2011), a topic of growing interest in the fMRI community. Traditionally, fMRI studies produce *forward inferences*: for a given task or paradigm, inference is made on the location of the brain response to the task. Reverse inference consists of using the pattern of brain activation to infer which task is most likely to have produced the data. If a neuroscientist has developed a new behavioural experiment (say, on emotion), he/she may indeed want to know whether their task engages the brain’s emotion processing system. In such case the researcher would want an estimate of the probability that the data arose from a population of emotion studies. Kang *et al.* (2014) show that classification based on the HPGRF model outperforms a naive classifier based on MKDA (MKDA + NBC). Reverse inference is naturally embedded into the spatial Bayesian latent factor model of Montagna *et al.* (2016), and classification performance outperforms MKDA + NBC on a meta-analysis of emotion and executive control studies. These results suggest that spatial models can capture information in the data that can not be learnt with a MUA.

For all these reasons, we believe that model-based methods have significant merits compared to kernel-based methods. However, there are several still open problem for model-based methods and CBMA in general. We discuss these open problems in the following section.

## 5 Open problems

Some aspects of CBMA are still being overlooked by both kernel-based and kernel-based methods. Arguably the most important is publication bias, which occurs when publicly accessible studies are not a representative subset of the total population of studies. A special case of publication bias is the so-called file drawer, that is, studies with significant findings are more likely to get published. If present, publication bias can affect the outcome of a meta-analysis and lead to false conclusions. While there is evidence for the existence of publication biases in fMRI (David *et al.*, 2013), there has been no attempt to quantify neither the extent of the phenomenon nor the effect these biases may have on meta-analysis estimates.

Another area of emerging interest is that of meta-regression. The use of study-specific characteristics as explanatory variables in a meta-analysis model is known as meta-regression (Greenland, 1994). Meta-regression is an important facet of meta-analysis, especially when there is appreciable heterogeneity between studies. As noted by Carp (2012), fMRI experiments exhibit strong heterogeneity,



hence it is essential to explore the effect that study characteristics have on the outcome of the analyses. Of both kernel and model-based approaches revised in Section 3, only the spatial Bayesian latent factor regression model of Montagna *et al.* (2016) can accommodate covariate information collected on the different studies. It is therefore a breakthrough contribution in neuroimaging meta-analysis research. By incorporating covariates in a meta-analysis on emotions, the authors observe, for example, that failing to adjust for multiple hypothesis testing results into a higher expected number of foci over the brain. The other CBMA methods described in Section 3 assume that any deviance from the true population effect is due to sampling error. Therefore, it is essential to extend these methods to account for meta-regression.

Another interesting avenue for future research consists in combining CBMA with image data from new fMRI studies. For example, there are no models that combine CBMA with IBMA when full 3D statistic images are available. Further, it is currently not possible to use meta-analysis data to improve estimation in small, underpowered group fMRI studies. A plausible direction is to build a point process model for single studies where the prior distribution for the study centers corresponds to the posterior intensity as obtained from a point process CBMA and IBMA model.

Often, additional information is available about the foci, such as the corresponding  $p$ -values or  $T$  scores. A convenient way to model these quantities is via mark processes, where these values become marks of the existing point patterns to improve estimation of the intensity function. Such an approach can enrich the inferences obtained from a meta-analysis by characterising the magnitude of activation at each voxel as opposed to modelling the location (and number) of the activations only, which is the question that current (model-based) methods address.

Finally, there is little work on functional connectivity for model-based methods. Functional connectivity refers to the dependency between one or more regions of the brain. In CBMA functional connectivity is implied by co-activation, that is, when two regions consistently report activations. In a recent work, Xue *et al.* (2014) use a multivariate Poisson model to induce correlation among the foci in several regions on interest. However, it would be interesting to extend the spatial models in order to capture these correlations as well.

## Acknowledgements

PS, TEN & TDJ were supported by NIH grant R01 NS075066-01A1; TEN was supported by a Wellcome Trust fellowship 100309/Z/12/Z.

## References

- Arminger, G. and Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, **63**(3), 271–300.
- Bartels, A. and Zeki, S. (2004). The neural correlates of maternal and romantic love. *Neuroimage*, **21**(3), 1155–1166.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), pp. 289–300.

- Button, K. S., Ioannidis, J. P. a., Mokrysz, C., Nosek, B. a., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, **14**(5), 365–76.
- Calhoun, V. D. and Pearlson, G. D. (2012). A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions. *NeuroImage*, **59**(1), 25–35.
- Carp, J. (2012). The secret lives of experiments: Methods reporting in the fmri literature. *NeuroImage*, **63**(1), 289 – 300.
- Caspers, J., Zilles, K., Beierle, C., Rottschy, C., and Eickhoff, S. B. (2014). A novel meta-analytic approach: Mining frequent co-activation patterns in neuroimaging databases. *NeuroImage*, **90**(0), 390 – 402.
- Cole, D. M., Smith, S. M., and Beckmann, C. F. (2010). Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Frontiers in systems neuroscience*, **4**(8).
- Costafreda, S., Brammer, M., David, A., and Fu, C. (2008). Predictors of amygdala activation during the processing of emotional stimuli. *Brain Research Reviews*, **58**(1), 57 – 70.
- Costafreda, S. G., David, A. S., and Brammer, M. J. (2009). A parametric approach to voxel-based meta-analysis. *NeuroImage*, **46**(1), 115 – 122.
- David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., Munafò, M. R., and Ioannidis, J. P. A. (2013). Potential Reporting Bias in fMRI Studies of the Brain. *PLoS ONE*, **8**(7).
- Delvecchio, G., Fossati, P., Boyer, P., Brambilla, P., Falkai, P., Gruber, O., Hietala, J., Lawrie, S. M., Martinot, J.-L., McIntosh, A. M., Meisenzahl, E., and Frangou, S. (2012). Common and distinct neural correlates of emotional processing in Bipolar Disorder and Major Depressive Disorder: a voxel-based meta-analysis of functional magnetic resonance imaging studies. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, **22**(2), 100–13.
- Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., and Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, **30**(9), 2907–2926.
- Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., and Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, **59**(3), 2349 – 2361.
- Etkin, A. and Wager, T. (2007). Functional Neuroimaging of Anxiety : A Meta-Analysis of Emotional Processing in PTSD , Social Anxiety Disorder , and Specific Phobia. *American Journal of Psychiatry*, **164**(10), 1476–1488.
- Farah, M. J. (2014). Brain images, babies, and bathwater: critiquing critiques of functional neuroimaging. *The Hastings Center report*, **44**(S2), S19–30.
- Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J. H., and Zamarripa, F. (1997). Functional volumes modeling: Theory and preliminary assessment. *Human Brain Mapping*, **5**(4), 306–311.

- Fox, P. T., Parsons, L. M., and Lancaster, J. L. (1998). Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Current Opinion in Neurobiology*, **8**(2), 178–187.
- Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: Theory. *NeuroImage*, **16**(2), 465 – 483.
- Fusar-Poli, P. (2012). Voxel-wise meta-analysis of fMRI studies in patients at clinical high risk for psychosis. *Journal of psychiatry & neuroscience : JPN*, **37**(2), 106–12.
- Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, **15**(4), 870–878.
- Greenland, S. (1994). Invited commentary: A critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, **140**(3), 290–296.
- Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. John Wiley & Sons, Hoboken.
- Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Academic Press.
- Huettel, S. A., Song, A. W., and McCarthy, G. (2009). *Functional Magnetic Resonance Imaging Second Edition*. Sinauer Associates, Inc, Massachussets.
- Illian, J., Penttinen, P., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.
- Kang, J., Johnson, T. D., Nichols, T. E., and Wager, T. D. (2011). Meta Analysis of Functional Neuroimaging Data via Bayesian Spatial Point Processes. *Journal of the American Statistical Association*, **106**(493), 124–134.
- Kang, J., Nichols, T. E., Wager, T. D., and Johnson, T. D. (2014). A bayesian hierarchical spatial point process model for multi-type neuroimaging meta-analysis. *The Annals of Applied Statistics*, **8**(3), 1561–1582.
- Kim, S.-G. and Ogawa, S. (2012). Biophysical and physiological origins of blood oxygenation level-dependent fMRI signals. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, **32**(7), 1188–206.
- Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., and Wager, T. D. (2008). Functional grouping and cortical and subcortical interactions in emotion: A meta-analysis of neuroimaging studies. *NeuroImage*, **42**(2), 998 – 1031.
- Konova, A. B., Moeller, S. J., and Goldstein, R. Z. (2013). Common and distinct neural targets of treatment: Changing brain function in substance addiction. *Neuroscience & Biobehavioral Reviews*, **37**(10), 2806–2817.
- Laird, A. R., Lancaster, J. J., and Fox, P. T. (2005). Brainmap: The social evolution of a human brain mapping database. *Neuroinformatics*, **3**(1), 65–77.
- Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, **16**(2), 538 – 550.

- Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, **23**(4), 439–464.
- Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton.
- Montagna, S., Tokdar, S. T., Neelon, B., and Dunson, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, **68**(4), 1064–73.
- Montagna, S., Wager, T., Feldman-Barrett, L., Johnson, T. D., and Nichols, T. E. (2016). Spatial Bayesian Latent Factor Regression Modeling of Coordinate-based Meta-analysis Data. *ArXiv e-prints*.
- Mumford, J. A. and Nichols, T. (2006). Modeling and inference of multisubject fmri data. *Engineering in Medicine and Biology Magazine, IEEE*, **25**(2), 42–51.
- Mumford, J. A. and Nichols, T. (2009). Simple group fmri modeling and inference. *NeuroImage*, **47**(4), 1469–1475.
- Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, **12**(5), 419–446.
- Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, **15**(1), 1–25.
- Nielsen, F. Å. and Hansen, L. K. (2002). Modeling of activation data in the BrainMap database: Detection of outliers. *Human Brain Mapping*, **15**(3), 146–156.
- Phelps, E. A. and LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, **48**(2), 175–187.
- Poldrack, R. (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding.
- Poldrack, R. A., Mumford, J. A., and Nichols, T. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press.
- Radua, J. and Mataix-Cols, D. (2009). Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. *The British journal of psychiatry : the journal of mental science*, **195**(5), 393–402.
- Radua, J. and Mataix-Cols, D. (2012). Meta-analytic methods for neuroimaging data explained. *Biology of mood & anxiety disorders*, **2**(1), 6.
- Radua, J., Mataix-Cols, D., Phillips, M. L., El-Hage, W., Kronhaus, D. M., Cardoner, N., and Surguladze, S. (2012). A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European Psychiatry*, **27**(8), 605–611.
- Radua, J., Rubia, K., Canales-Rodríguez, E. J., Pomarol-Clotet, E., Fusar-Poli, P., and Mataix-Cols, D. (2014). Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies. *Frontiers in psychiatry*, **5**(February), 13.

- Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R., Kahn, R., and Ramsey, N. (2007). Test-retest reliability of fmri activation during prosaccades and antisaccades. *NeuroImage*, **36**(3), 532 – 542.
- Richlan, F., Kronbichler, M., and Wimmer, H. (2011). Meta-analyzing brain dysfunctions in dyslexic children and adults. *NeuroImage*, **56**(3), 1735–42.
- Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009). Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, **45**(3), 810 – 823.
- Shermer, M. (2008). Why You Should Be Skeptical of Brain Scans. *Scientific American Mind*, **19**(5), 66–71.
- Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme, Stuttgart.
- Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., and Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage*, **35**(1), 105 – 120.
- Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: Method and validation. *NeuroImage*, **16**(3, Part A), 765 – 780.
- Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., and Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, **33**(1), 1–13.
- Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, **4**(3), 274–290.
- Wager, T., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Social Cognitive and Affective Neuroscience*, **2**(2), 150–158.
- Wager, T. D., Phan, K., Liberzon, I., and Taylor, S. F. (2003). Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *NeuroImage*, **19**(3), 513–531.
- Wager, T. D., Jonides, J., and Reading, S. (2004). Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage*, **22**(4), 1679 – 1693.
- Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**(2), 251–267.
- Xue, W., Kang, J., Bowman, F. D., Wager, T. D., and Guo, J. (2014). Identifying functional co-activation patterns in neuroimaging studies via poisson graphical models. *Biometrics*, **70**(4), 812–822.
- Yue, Y. and Speckman, P. L. (2010). Nonstationary spatial gaussian markov random fields. *Journal of Computational and Graphical Statistics*, **19**(1), 96–116.

Yue, Y. R., Lindquist, M. A., and Loh, J. M. (2012). Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *The Annals of Applied Statistics*, **6**(2), 697–718.